Annotation of the Corpus of the Saeima with Multilingual Standards

Roberts Darģis, Ilze Auziņa, Uldis Bojārs, Pēteris Paikens, Artūrs Znotiņš
Motivation

• Current comparative research of parliamentary debate is not sufficiently facilitated

• Augmenting such corpora with extra layers according to multilingual standards would help

• Annotation Layers
  • English translation
  • Named entities
  • Morphological and syntactical annotations

• Available datasets
  • Linked Data
  • Universal Dependencies
  • Bonito corpus browser
The Corpus of the Saeima

- Saeima has been the name of the parliament of the Republic of Latvia since 1922
- The Saeima is composed of 100 members of parliament and it is elected for a term of four years
- The Corpus of the Saeima includes transcriptions of parliamentary debate from 7 parliamentary terms (5th–12th), covering years 1993–2017
- The transcriptions of the Corpus of the Saeima contain 38 million tokens, 497 thousand utterances and 468 speakers
Morphological and Syntactical Annotations

- The annotations contain lemma, part of speech, morphological features and syntactic dependencies according to the Universal Dependencies standard format.
- Texts are automatically tokenized, lemmatized and morphologically analyzed and tagged using CMM based tagger.
- Syntactic dependencies are inferred by neural transition-based dependency parser trained on Latvian Universal Dependencies corpus version 2.1.
Bonito corpus browser (NoSketch engine)

- The interface provides powerful corpus query system. Query can include words, lemmas, morphological tags and meta data

http://dati.saeima.korpuss.lv/nosketch
Universal Dependencies (CoNLL-U)

- Automatic tokenization, morphological and syntactic annotations are published in CoNLL-U data format

```
# newdoc id = 2016_03_31_355.txt_seq17
# newpar id = 2016_03_31_355.txt_seq17-p1
# sent_id = 2016_03_31_355.txt_seq17-p1s1
# text = Turpinām ar iesniegtajām izmaiņām Saeimas Prezidīja apstiprinātajā sēdes darba kārtībā.

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Turpinām</td>
<td>turpināt</td>
<td>_</td>
<td>vmnipt31pan</td>
<td>_</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>ar</td>
<td>_</td>
<td>_</td>
<td>sppd</td>
<td>_</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>iesniegtajām</td>
<td>iesniegt</td>
<td>_</td>
<td>vmnpdfpdpysyp</td>
<td>_</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>izmaiņām</td>
<td>izmaiņa</td>
<td>_</td>
<td>ncfpd4</td>
<td>_</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Saeimas</td>
<td>saeima</td>
<td>_</td>
<td>ncfsg4</td>
<td>_</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>Prezidīja</td>
<td>prezidījs</td>
<td>_</td>
<td>ncmgs1</td>
<td>_</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>apstiprinātajā</td>
<td>apstiprināt</td>
<td>_</td>
<td>vmnpdfsllpysyp</td>
<td>_</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>sēdes</td>
<td>sēde</td>
<td>_</td>
<td>ncfsg5</td>
<td>_</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>darba</td>
<td>darbs</td>
<td>_</td>
<td>ncmgs1</td>
<td>_</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>kārtībā</td>
<td>kārtība</td>
<td>_</td>
<td>ncfls4</td>
<td>_</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>.</td>
<td>_</td>
<td>_</td>
<td>zs</td>
<td>_</td>
<td>1</td>
</tr>
</tbody>
</table>
```
Machine Translation to English

• The speeches from Latvian are translated to English using a neural machine translation system
• The unreviewed machine-generated translation is provided for quantitative analysis and to aid searchability and understanding for international researchers
• The text quality of automated translation is lacking, so for qualitative analysis a professional translator should be used
Named Entities

• We used the structured Wikidata information extracts as the entity knowledge base. The Wikidata entity alias information is extended with Latvian morphological inflections and automatically generated variants for people and organization names.

• In the Corpus of the Saeima we identified 393 thousand mentions of 3 thousand unique entities. 165 thousand out of 497 thousand utterances contained entity mentions.
LinkedSaeima I – structure

- Structured information about parliamentary debates is represented using Resource Description Framework (RDF), according to the Linked Data principles

- The types of objects in the LinkedSaeima dataset are:
  - Meeting – a top-level concept representing one parliament meeting (a plenary) usually consisting of multiple Speeches
  - Speech – an individual speech given at a Meeting by a particular Speaker in some Role
  - Speaker – a person giving a speech
  - Role – a role (e.g. Prime Minister) which the person represented when giving a Speech
LinkedSaeima II – interfaces

Screenshot of the LinkedSaeima triple pattern fragments server

Screenshot of a LinkedSaeima entity in LodView

LinkedSaeima index page - http://dati.saeima.korpuss.lv
LinkedSaeima III – innovation

• Main innovation of this dataset, relative to the LinkedEP project:
  • Addition of named entity information, pointing to corresponding Wikidata URI identifiers
  • “Materialization” of speaker Roles, by giving them URI identifiers and linking them to Wikidata URI identifiers
  • Manually linking speakers to Wikidata URI, to make it easier to conduct a inter-corpora research
Conclusions

• The new annotation levels and its Linked Data representation will widen the applications for *The Corpus of the Saeima*

• Future work includes improvements to annotation tools, and extending the LinkedSaeima dataset

• We’d like to call upon this research community to pursue open, common NLP data standards to enable multilingual comparative research
Thank you!

Resource and description: http://dati.saeima.korpuss.lv